Sequencing By Ligation Variation with Endonuclease V and Deoxyinosine and SAWTooth – The Sequencing Analysis Workbench Tool

Antoine Ho[1], Maurice Murphy[2,3], Susan Wilson[2,3], Susan R. Atlas[2,3,4], Jeremy S. Edwards[1,2,5]

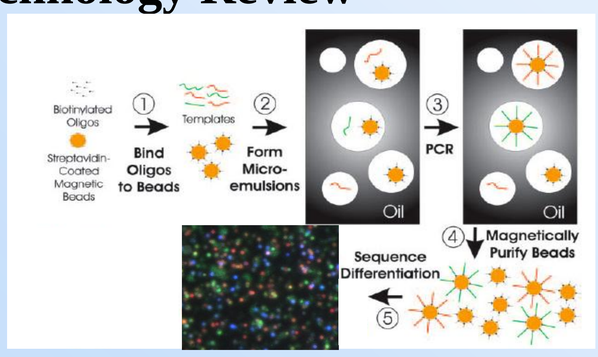
[1]UNM Department of Molecular Genetics and Microbiology, [2]UNM Cancer Center, [3]UNM Center for Advanced Research Computing, [4]UNM Department of Physics and Astronomy, [5]Department of Chemical and Nuclear Engineering, University of New Mexico 87131

Abstract

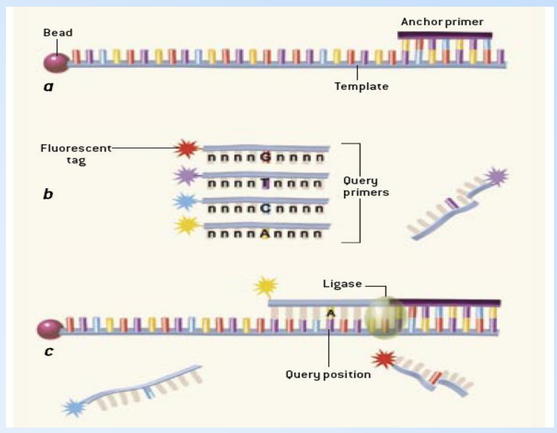
Sequencing-by-ligation (SBL) is one of several next-generation sequencing methods that has been developed for massive sequencing of DNA immobilized on arrayed beads (or other clonal amplicons). SBL has the advantage of being easy to implement and accessible to all because it can be performed with off-the-shelf reagents. However, SBL has the limitation of very short read lengths. To overcome the read length limitation, research groups have developed complex library preparation processes, which can be time-consuming, difficult, and result in low complexity libraries. Herein we describe a variation on traditional SBL protocols that extends the number of sequential bases that can be sequenced by using Endonuclease V to nick a query primer, thus leaving a ligatable end extended into the unknown sequence for further SBL cycles.

Additionally, virtually all next-generation sequencing platforms generate giga-base-pairs of data per run, often in the form of mate-paired short–reads. We anticipate the daily need to sequence, and subsequently align (map) to a reference genome, several billion mate-pair reads, or single sequence reads, in whole-genome sequencing of human samples. These reads may need to be aligned to a large reference genome, itself comprising several giga-base-pairs. An efficient algorithm to perform this mapping is essential given these large dataset sizes. Here we present the SawTooth suite of software applications whose core functionality is the efficient mapping of short-read sequencing data to a reference genome. SawTooth also implements several ancillary applications for validation and statistical analysis of mapping results.

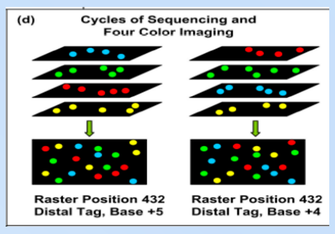
Current Polonator Technology Review



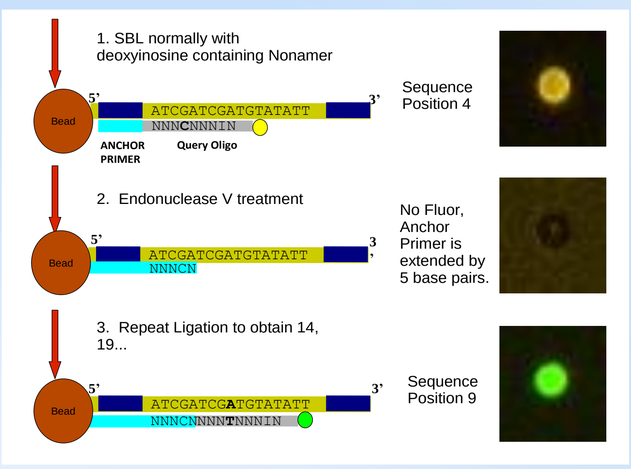
Polony Sequencing utilizes a fixed Bead array and Sequencing By Ligation (SBL) to obtain sequence information. Biotinylated template DNA is isolated in a PCR solution and added to streptavidin coated beads. Through a process of emulsion PCR, the beads become clonally covered with multiple copies of a single template strand. In SBL, an anchor primer is hybridized to a known region of the template DNA, usually a linker or adaptor that has been ligated onto a fragment of unknown DNA sequence. Then using a series of degenerate query oligos that have fluorophores, the DNA adjacent to the known region is sequenced through a series of hybridizing an anchor primer, ligation of a query oligo, then denaturing the DNA and clearing all signal and repeating.



Images captured in four fluorescent channels, one corresponding to each base pair, serve as data. On every frame, every bead's coordinates are recorded as well as the signal the bead gave for a base pair position. After multiple cycles of biochemistry and imaging yields a sequence for every bead, which is then used as the raw sequence for alignment.



Cyclic Sequencing By Ligation with Endonuclease V

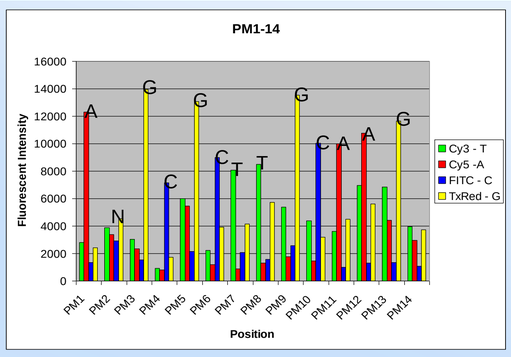


Cyclic SBL involves using Endonuclease V to recognize an incorporated deoxyinosine site and clip the DNA.

This would result in a fragment of the query primer attached to the anchor primer, allowing subsequent ligations to sequence further out.

Cyclic SBL Results

Results indicate that it is possible to proceed through three iterations of cyclic SBL, or in our particular test case, 13 base pairs in a single direction. With a flanking adaptor and the proper Type IIs restriction enzyme, a read length of >20 bases, or >40 mate paired bases per tag would be possible.

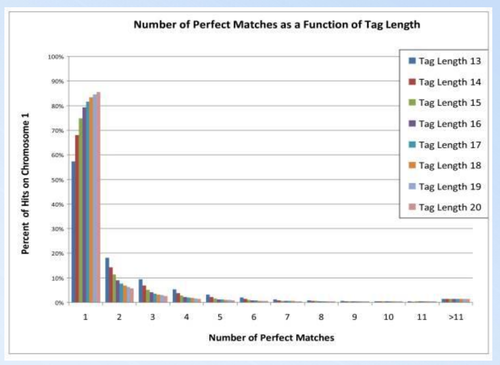


Test Template (Anchor Regions Underlined):
5' TCT ATG GGC AGT CGG TGA TAN GCG CTT GCA AGA GAA TGA GGA AAA CGA AGA 3'
Fluorescent Intensity plotted versus position for each channel.

Proof-of-Concept Coverage Simulation of Chromosome

To demonstrate the feasibility of a cSBL approach to genome sequencing and calculate gains in using cSBL over traditional SBL methods, we utilized the SawTooth resequencing code developed at the University of New Mexico (M. Murphy *et al.*, to be submitted, 2011). Human genome coverage was simulated using mate-paired data ranging from twenty-six bases to (limit of traditional SBL) to forty bases (theoretical gain from cSBL implementation).

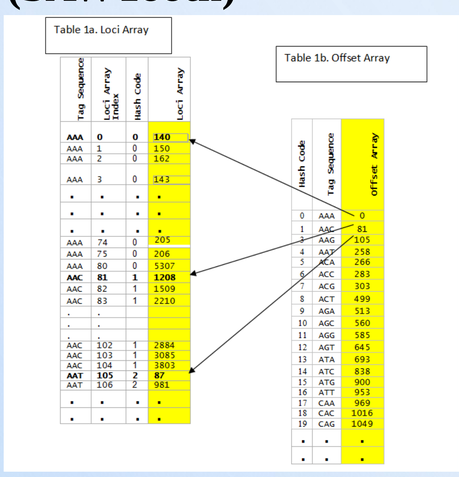
A set of simulated mate-paired tags, each separated by a range of 300-700 bases, was created, ranging in size from 13 paired tags to 20 paired tags. A sufficient number of tags were computationally generated to simulate 10 × coverage. The tags were all generated from chromosome 1, mapped back to the entire genome, and calculations of chromosome 1 coverage were performed.



Next, we performed an analysis of how many times each tag mapped to the genome. One of the more significant benefits gained by increasing tag length from 13 to 20 bases is that far fewer tags must be discarded because they do not map uniquely. At a tag length of 13 bases, only 57.2% of the tags are used, compared to 85.6% at a tag length of 20, thus effectively increasing throughput.

Sequencing Analysis Workbench Tool (SAWTooth)

Next-generation sequencing platforms generate gigabytes of data per run, often in the form of mate-paired reads. This requires the analysis and mapping of several billion mate-paired reads when used for whole-genome sequencing. An efficient algorithm to perform this mapping is essential given these large dataset sizes. SAWTooth was developed for the purpose of efficient mapping of short-read sequencing data to a reference genome, outperforming other popular codes used in genome alignment by ~ 100-fold or more.



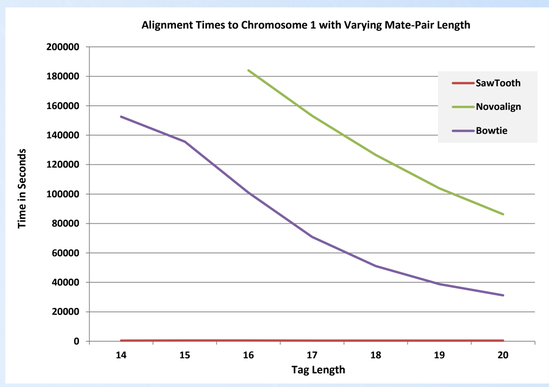
An example of loci data structures using 3-mers.

All fast contemporary mapping algorithms rely on indexes. These auxiliary data structures facilitate mapping sequences to a reference genome. These indexes generally fall into two broad categories – suffix trees and hash indexes.

Traditionally, the construction and use of suffix-trees imposed prohibitive memory requirements, though in recent years, innovations in the field of compressed text indexes have rendered suffix-based methods feasible for whole-genome indexing, though still not optimal.

SAWTooth utilizes hash indexing, a well-known referencing data structure which allows key-based data retrieval in constant, O(1) time, making it the fastest of all data retrieval structures. In principle, there are some limitations of general hash indexes that may limit their performance or impair their usefulness. Keys are not ordered, so sorted lists and range searches are not intrinsic operations on the data structures. Also, hash function may generate the same hash for multiple keys, i.e. a collision. Resolving collisions require extra processing and access to the original keys within the index.

However, the special nature of genomic data and our specialized purpose of mapping mate-paired reads to a reference genome, allow us to create hash indexes that are free from these limitations. In SawTooth, the hash key is the sequence tag, and the data to be retrieved is an exhaustive list of loci where the reads map in the reference genome.



Tag Size (bp)	SawTooth (sec.)	Novoalign (sec.)	Bowtie (Seconds)	Speedup Relative to Novoalign	Speedup Relative to Bowtie
14	463		152588	0	329
15	558		135556	0	243
16	519	184003	100919	355	194
17	479	153142	70849	319	148
18	448	126530	51069	282	114
19	428	103814	38754	242	90
20	409	86302	31236	211	76

Benchmark Test: Search times to map 10M paired-end tags of various sizes 14-20.
Tags were artificially created from chromosome 1 to simulate speed of perfect matches, and later matched back to entire genome.

References

Collins, F.S., Morgan, M. and Patrinos, A. (2003) The Human Genome Project: Lessons from Large-Scale Biology. Science, 300, 286-290.

Drmanac, R. (2009). Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327, 78.

Antoine Ho, Maurice Murphy, Susan Wilson, Susan R Atlas and Jeremy S Edwards. Sequencing by ligation variation with endonuclease V digestion and deoxyinosine-containing query oligonucleotides. BMC Genomics 12:598.

Kato, K. (2009) Impact of the Next Generation DNA Sequencers. International Journal of Clinical and Experimental Medicine, 2, 193-202.

Metzker, M., (2010) Sequencing Technologies – The Next Generation. Nature Reviews Genetics, 11, 31-46.

Maurice Murphy, Susan Wilson, Antoine Ho, Jeremy S. Edwards, Susan R. Atlas. SAWTooth: Sequencing Analysis Workbench Tool. (In preparation)

Acknowledgements

Funding is provided by the National Institute of Health [R21 HG004350/564251] and the National Science Foundation, IGERT INCBN Graduate Fellowship [DGE-0549500].

We thank the UNM Center for Advanced Research Computing and the UNM Cancer Center Shared Resource for Bioinformatics and Computational Biology for computational resources in support of this work.